Error Bounds for a Diffusion Model-Based Drift Estimator

Anonymous authors
Paper under double-blind review

Abstract

Parameter estimation in stochastic differential equations is a classical statistical problem of much importance in many scientific fields. Recent work of Tapia Costa et al. (2026) introduced a novel technique for estimating the drift when the diffusion parameter is known, using discrete samples from multiple trajectories. Their method treats drift estimation as a denoising problem, and leverages tools from (conditional) score-matching diffusion models. Although their experiments showed promising results across different drift classes, the question of theoretical guarantees for their estimator was left unanswered. In this note, we address this gap by exploiting techniques from diffusion model theory. More concretely, we derive an explicit risk bound for the time-averaged mean-squared error of said drift estimator. Our bound decomposes the risk into the (i) Euler-Maruyama discretization, (ii) score/denoiser approximation, (iii) noise initialization, and (iv) sampling variance, revealing the trade-offs between the different hyperparameters and sources of error in the estimator.

1 Introduction

Stochastic differential equations (SDEs) are mathematical models playing a fundamental role for modeling dynamical systems that are subject to deterministic trends and randomness, with important applications across physics, biology, finance, engineering, and the geosciences. Hence, reliable parameter estimation is essential both for scientific understanding and for prediction and control. In particular, estimating the drift –that is, the deterministic part– of an SDE from discretely observed sample paths is a classical problem in statistics (Kutoyants, 2004).

Recently, Tapia Costa et al. (2026) proposed a novel approach for this task. Instead of learning the drift directly via regression from the increments, which is the classical approach, they introduce a conditional denoising diffusion model for learning noisy one-step transitions and then recover the drift from this denoiser. Their approach connects drift estimation with the score-matching and denoising ideas that underlie modern diffusion models, and suggests a possible way to stabilize learning in regimes where direct regression may be difficult or unstable.

Although Tapia Costa et al. (2026) show empirically that their method is competitive across drift classes, they do not provide a theoretical analysis of their estimator. This note complements their contribution by providing an explicit bound for the time-averaged mean-squared error of their drift estimator. The resulting decomposition separates the contribution of the Euler–Maruyama discretization, the score approximation error, noise initialization error, and the Monte Carlo estimation error. In this way, the result makes precise which parts of the method are responsible for the final estimation error, and how the main hyperparameters enter the trade-off. Remarkably, our bound captures the nonlinear relation between sampling frequency and mean-squared error observed experimentally by Tapia Costa et al. (2026).

The paper is structured as follows: in Section 2 we introduce the prerequisites for our main results, namely, score-matching diffusion models (SDMs) and the SDM-based drift estimator of Tapia Costa et al. (2026). In Section 3 we obtain error bounds for the time-averaged mean-squared error of the estimator and discuss their implications. Finally, the limitations of our approach and avenues for future work are discussed in Section 4.

1.1 Related work

The statistical literature on drift estimation for diffusion processes is extensive; see Kutoyants (2004) for general background and Section 2.1 of Tapia Costa et al. (2026) for a recent overview. Many works formulate drift estimation as a nonparametric regression problem (Comte et al., 2007; Comte & Genon-Catalot, 2020; Zhao et al., 2020; Denis et al., 2021), and more recent papers have considered neural-network-based regression (Oga & Koike, 2024; Zhao et al., 2025).

In our setting, the training data are discrete observations from a finite set of i.i.d. trajectories from the SDE, observed at regular time intervals; see Subsection 2.2 for the mathematical formulation of the problem. There is also a growing literature that aims to learn stochastic dynamics, but using uncorrelated samples from the marginals (Neklyudov et al., 2023; Guan et al., 2024; Lavenant et al., 2024).

Our analysis of the diffusion-model-based estimator of Tapia Costa et al. (2026) is informed by techniques from the convergence theory of SDMs such as Chen et al. (2023); Mbacke & Rivasplata (2024); Gao et al. (2025a) or Pfarr et al. (2026). See Section 6 of Tang & Zhao (2025) for a survey on different approaches to diffusion model theory. We also note several recent diffusion model-based approaches that are adjacent to Tapia Costa et al. (2026) in spirit. These include conditional diffusion methods for sampling SDE trajectories (Liu et al., 2025; Gao et al., 2025b). However, these methods do not directly target drift estimation, and their convergence theory is largely unexplored.

2 Preliminaries

2.1 Introduction to Score-Matching Diffusion Models (SDMs)

SDMs are a family of generative models that aim to generate new samples from an unknown distribution by learning to reverse a noise-injection procedure. Starting from the target distribution $X_0 \sim p_0$, they gradually inject randomness until they end up with (almost) pure noise

$$X_0 \to X_1 \to ... \to X_{\mathcal{T}} \sim p_{\mathcal{T}} \approx p_{\text{noise}}. \tag{1}$$

This forward process is straightforward. The crux of diffusion models is to learn how to reverse it; that is, starting from $X_{\mathcal{T}} \sim p_{\text{noise}}$, how to generate a sample from p_0 by inverting the forward process

$$X_{\mathcal{T}} \to X_{\mathcal{T}-1} \to ... \to X_0 \sim p_0. \tag{2}$$

Formally, the forward process can be understood as a discretization of a (Markov) diffusion process, while the second step amounts to learning its corresponding time-reversed process. The problem can be thus expressed naturally in the language of stochastic differential equations (SDEs). The presentation below mainly follows Tang & Zhao (2025).

Consider the continuous-time process $(X_\tau)_{\tau\in[0,\mathcal{T}]}$ described by the following SDE:

$$dX_\tau = f(\tau, X_\tau)d\tau + g(\tau)dB_\tau, \quad X_0 \sim p_0, \tag{3}$$

where $f : \mathbb{R}_+ \times \mathbb{R}^D \to \mathbb{R}^D$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ are respectively the drift and noise parameters, and $(B_\tau)_{\tau\geq 0}$ is D-dimensional Brownian motion (Le Gall, 2016). Note that, from a SDM perspective, f and g are chosen by the user: different diffusion models correspond to different choices of f and g.

Now consider the time-reversal of the process (3), i.e., $(\widetilde{X}_\tau)_{\tau\in[0,\mathcal{T}]} := (X_{\mathcal{T}-\tau})_{\tau\in[0,\mathcal{T}]}$. Under mild assumptions on f and g, a theorem of Anderson (1982) shows that $(\widetilde{X}_\tau)_{\tau\in[0,\mathcal{T}]}$ satisfies the SDE

$$d\widetilde{X}_\tau = \left[-f(\mathcal{T}-\tau, \widetilde{X}_\tau) + g(\mathcal{T}-\tau)^2\nabla \log p_{\mathcal{T}-\tau}(\widetilde{X}_\tau)\right] d\tau + g(\mathcal{T}-\tau)dB_\tau, \quad \widetilde{X}_0 \sim p_{\mathcal{T}}, \tag{4}$$

where p_τ is the probability density of X_τ conditional on X_0.

It is worth noticing that the diffusion term in the SDE for $\widetilde{X}_\tau$ is given by the same function as the corresponding term in the SDE for X_τ, namely g, which is evident by looking at equations (4) and (3). By

contrast, the drift term in (4) uses not only the drift function f in (3) but also additionally uses the diffusion function g and the score function $\nabla \log p_\tau(X_\tau)$.

At this point, if we could sample from $p_{\mathcal{T}}$ and solve (4), we would already be able to generate samples from p_0 by running the reverse process. However, several obstacles remain in our way.

First, except for very specific situations, $p_{\mathcal{T}}$ still depends on p_0 via X_0, so generally we are not able to sample from $p_{\mathcal{T}}$. A widely-used solution is to substitute $p_{\mathcal{T}}$ with some distribution p_{noise} from which it is easy to sample, and which is still close enough to $p_{\mathcal{T}}$. Second, the score function, $\nabla \log p_\tau(X_\tau)$, is unknown. In fact, the core idea behind SDMs is to estimate $\nabla \log p_\tau(X_\tau)$ by training a neural network $s^\theta_\tau(x)$ using score-matching (Hyvärinen & Dayan, 2005). See Section 4.2 on Lai et al. (2025) or Section 4 in Tang & Zhao (2025) for a discussion of several techniques. Finally, after these approximations, we can solve (4) using SDE discretization schemes such as the Euler-Maruyama method (Kloeden & Platen, 1992).

2.1.1 Example: Variance preserving (VP) SDEs

VP SDEs were introduced by Song et al. (2021) as the continuous limit of denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020). DDPMs use N noise scales $\beta_1 < \cdots < \beta_N$ and run the (forward) process

$$x_i = \sqrt{1-\beta_i}x_{i-1} + \sqrt{\beta_i}z_i, \quad 1 \le i \le N, \tag{5}$$

where z_i are i.i.d. samples from $\mathcal{N}(0, I_D)$. It can be shown (see Section 3(c) of Tang & Zhao (2025) for the details) that this scheme corresponds to the continuous-time SDE

$$dX_\tau = -\frac{1}{2}\beta(\tau)X_\tau d\tau + \sqrt{\beta(\tau)}dB_\tau, \quad 0 \le \tau \le \mathcal{T}, \tag{6}$$

which is a particular case of (3) with the choices $f(\tau, x) = -\frac{1}{2}\beta(\tau)x$ and $g(\tau) = \sqrt{\beta(\tau)}$. By solving (6) we get the distribution

$$p_\tau(\cdot|X_0 = x) = \mathcal{N}\left(e^{-\frac{1}{2}\int_0^\tau \beta(s)ds}x, \left(1 - e^{-\int_0^\tau \beta(s)ds}\right) I_D\right). \tag{7}$$

For common choices of $\beta(\tau)$ and $\mathcal{T}$ large enough, it is thus reasonable to choose $p_{\text{noise}} = \mathcal{N}(0, I_D)$, which finally results in the following approximate reverse SDE:

$$d\widetilde{X}_\tau = \left[\frac{1}{2}\beta(\mathcal{T}-\tau)\widetilde{X}_\tau + \beta(\mathcal{T}-\tau)s^\theta_{\mathcal{T}-\tau}(\widetilde{X}_\tau)\right]d\tau + \sqrt{\beta(\mathcal{T}-\tau)}dB_\tau, \quad \widetilde{X}_0 \sim \mathcal{N}(0, I_D). \tag{8}$$

While Tapia Costa et al. (2026) used the specific schedule $\beta(\tau) = \gamma_0 + \tau(\gamma_1 - \gamma_0)$, for some $\gamma_1 > \gamma_0 > 0$, the results we present here are valid for arbitrary $\beta(\tau)$.

2.2 Estimating the drift of an SDE

Suppose we are interested in estimating the (time-homogeneous) drift function, $\mu : \mathbb{R}^D \to \mathbb{R}^D$, of the following continuous-time SDE:

$$dY_t = \mu(Y_t)dt + \sigma dB_t, \quad t \in [0, T+\Delta], \tag{9}$$

where the diffusion parameter $\sigma > 0$ is known. For this purpose, we are given a dataset $\mathcal{D}_{\text{obs}}$ consisting of i.i.d. trajectories from (9), say I of them, each trajectory with J observations obtained at times $t_j := j\Delta$ for $j = 0, \dots, J$, where $\Delta = T/J$ is the frequency at which we gather observations. That is,

$$\mathcal{D}_{\text{obs}} = \left\{Y^{(i)}_{t_0}, \dots, Y^{(i)}_{t_J}\right\}, \quad i = 1, \dots, I. \tag{10}$$

Remark 2.1. We define the SDE (9) on $[0, T+\Delta]$ for notational convenience and to make the bound in Theorem 3.5 cleaner. The SDE can be defined in $[0, T]$ at the cost of dividing by $T - \Delta$ and integrating up to $T - \Delta$ in the time-averaged MSE error (22).

The classical approach to the drift estimation problem is regression over the increments. Let $Z_{t_j} := Y_{t_j} - Y_{t_{j-1}}$, for $j = 1, \dots, J$. The solution to (9) on $[t_{j-1}, t_j)$ is of the form

$$Z_{t_j} = \int_{t_{j-1}}^{t_j} \mu(Y_s)ds + \sigma\sqrt{\Delta}\omega, \quad \omega \sim \mathcal{N}(0, I_D). \tag{11}$$

Observe that, due to the Markov property of $(Y_t)_{t\in[0,T]}$ and the time-homogeneity of μ, the random variables $Z_{t_j} \,|\, Y_{t_{j-1}}$ are identically distributed for every $j = 1, \dots, J$. Now, given the frequency Δ, an Euler–Maruyama (EM) approximation to (11) yields

$$Z_{t_j} \approx \widehat{Z}_{t_j} := \mu(Y_{t_{j-1}})\Delta + \sigma\sqrt{\Delta}\omega, \quad \omega \sim \mathcal{N}(0, I_D), \tag{12}$$

with $\widehat{Z}_{t_j} \,|\, (Y_{t_{j-1}} = y) \sim \mathcal{N}\left(\mu(y)\Delta, \sigma^2\Delta I_D\right)$, and $\mathbb{E}\left[\widehat{Z}_{t_j} \,|\, Y_{t_{j-1}} = y\right] = \mu(y)\Delta$. This suggests learning μ by regression:

$$\min_{\theta} \sum_{i\in I, j\in J} \left\| D_\theta\big(Y^{(i)}_{t_{j-1}}\big) - Z^{(i)}_{t_j} \right\|_2^2, \tag{13}$$

where D_θ is a fixed class of functions. However, Tapia Costa et al. (2026) argue that this approach can suffer from high variance and the curse of dimensionality. Inspired by approaches that regularize regression problems with input noise, they propose using a conditional diffusion model to estimate the SDE drift.

2.2.1 Drift estimation with diffusion models

Their key observation is that, since the increments Z_{t_j} can be approximated as noisy versions of the signal $\mu(Y_{t_{j-1}})\Delta$, one can treat drift estimation as a denoising problem. Concretely, Tapia Costa et al. (2026) apply the VP SDE framework of Section 2.1.1 to p_0, the density of $Z_{t_j} \,|\, Y_{t_{j-1}}$, and then extract the drift from the learned denoiser. As mentioned above, the law of $Z_{t_j} \,|\, Y_{t_{j-1}}$ is the same for every t_j, so we just write $Z \,|\, Y$ from now on.

We now apply the VP forward process to the increments Z, treating Y as a fixed conditioning variable. That is, Z plays the role of X_0 in (6). For a given noise level $\tau > 0$, the forward kernel produces the noisy version

$$X_\tau = \alpha_\tau Z + \sigma_\tau \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_D), \tag{14}$$

where $\alpha_\tau = e^{-\frac{1}{2}\int_0^\tau \beta(s)\,ds}$ and $\sigma_\tau^2 = 1 - \alpha_\tau^2$ are exactly as in (7). The conditional marginal at noise level τ given $Y = y$ is thus

$$p_\tau(x_\tau \mid Y = y) = \int_{\mathbb{R}^D} \mathcal{N}\big(x_\tau;\ \alpha_\tau z,\ \sigma_\tau^2 I_D\big)\, p_0(z \mid y)\, dz. \tag{15}$$

A neural denoiser $D_\theta(\tau, x_\tau, y)$ is then trained to predict the increment Z from the noisy version X_τ and the conditioning state Y, by minimizing the conditional denoising loss

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau,Y,Z,X_\tau} \left\| D_\theta(\tau, X_\tau, Y) - Z \right\|_2^2, \tag{16}$$

where the expectation is over $\tau \sim \mathcal{U}[\varepsilon, \mathcal{T}]$, (Y, Z) drawn from the pooled training pairs, and X_τ sampled via equation (14). The population minimizer of this conditional denoising loss is $\mathbb{E}_{p_0}[Z \mid X_\tau, Y]$.

Remark 2.2 (Connection to score-matching). The loss (16) is equivalent to the conditional score-matching loss up to reparametrization. Indeed, the conditional score of the VP SDE is

$$\nabla_{X_\tau} \log p_\tau(X_\tau \mid Y) = -\frac{X_\tau}{\sigma_\tau^2} + \frac{\alpha_\tau}{\sigma_\tau^2}\mathbb{E}_{p_0}[Z \mid X_\tau, Y],$$

so learning the conditional score is equivalent to learning the denoiser $\mathbb{E}_{p_0}[Z \mid X_\tau, Y]$.

The trained denoiser $D_{\theta^*}(\tau, x_\tau, y) \approx \mathbb{E}_{p_0}[Z \mid X_\tau = x_\tau, Y = y]$ approximates the posterior mean of the true increments. To obtain a closed-form relation between the denoiser and the drift $\mu(y)$, Tapia Costa et al. (2026) exploit the EM approximation (12):

$$Z \mid (Y = y) \ \approx\ \hat{p}_{\text{data}}(\cdot \mid y) := \mathcal{N}\big(\mu(y)\Delta,\ \sigma^2\Delta\, I_D\big).$$

Under this Gaussian approximation, one obtains the conditional mean

$$\mathbb{E}_{\hat{p}_{\text{data}}}[Z \mid X_\tau = x_\tau,\, Y = y] = \frac{\sigma_\tau^2 \Delta}{\sigma_\tau^2 + \alpha_\tau^2 \sigma^2 \Delta} \left(\mu(y) + \frac{\alpha_\tau \sigma^2}{\sigma_\tau^2}\, x_\tau \right). \tag{17}$$

Finally, solving equation (17) for $\mu(y)$ yields

$$\mu(y) = a(\tau)\, x_\tau + b_\Delta(\tau)\, \mathbb{E}_{\hat{p}_{\text{data}}}[Z \mid X_\tau = x_\tau,\, Y = y], \tag{18}$$

where

$$a(\tau) := -\frac{\alpha_\tau \sigma^2}{\sigma_\tau^2}, \qquad b_\Delta(\tau) := \frac{\sigma_\tau^2 + \alpha_\tau^2 \sigma^2 \Delta}{\sigma_\tau^2 \Delta}. \tag{19}$$

Observe that the identity (18) is only exact under the EM Gaussian approximation. This distinction is central to the error analysis in Theorem 3.5.

Replacing the conditional expectation in (18) by the learned denoiser D_{θ^*}, we obtain the single-sample plug-in drift estimator

$$\hat{\mu}(\tau, x_\tau, y) := a(\tau)\, x_\tau + b_\Delta(\tau)\, D_{\theta^*}(\tau, x_\tau, y). \tag{20}$$

In practice, the estimator is averaged over K i.i.d. samples $G^{(1)}, \ldots, G^{(K)} \sim \mathcal{N}(0, I_D)$:

$$\bar{\mu}_K(\tau, y) := \frac{1}{K} \sum_{k=1}^{K} \hat{\mu}\big(\tau,\, G^{(k)},\, y\big). \tag{21}$$

Hence the training data is only used to learn the denoiser D_{θ^*}, while at test time the estimator can, in principle, be evaluated at any state y for which the denoiser is defined.

Our theoretical guarantees below are stated for the time-averaged MSE error

$$\frac{1}{T} \int_0^T \mathbb{E} \big\| \mu(Y_t) - \bar{\mu}_K(\tau, Y_t) \big\|_2^2, \tag{22}$$

where the expectation is taken with respect to Y_t and the samples $G^{(k)} \sim \mathcal{N}(0, I_D)$ in $\bar{\mu}_K(\tau, y)$. This is the natural error metric for drift estimation; see (Comte & Genon-Catalot, 2020; Oga & Koike, 2024; Tapia Costa et al., 2026).

3 Main results

We now introduce assumptions on the SDE in (9) and the learned denoiser, D_{θ^*}, that will allow us to obtain our main result, a time-averaged mean-squared error bound for the estimator (21).

Assumption 3.1 (Lipschitz drift)**.** There is a constant $L > 0$ such that, for every $x, y \in \mathbb{R}^D$,

$$\|\mu(x) - \mu(y)\|_2 \le L \|x - y\|_2.$$

Assumption 3.2 (Finite second moment)**.** Let $(Y_t)_{t \in [0, T+\Delta]}$ be a solution to (9). Then for every $t \in [0, T + \Delta]$, we have

$$\mathfrak{m}_\mu(t) := \mathbb{E} \|\mu(Y_t)\|_2^2 < \infty.$$

Assumption 3.3 (L^2-accurate score-matching)**.** For any $\tau > 0$ there is $\varepsilon_{\text{score}}(\tau) > 0$ such that

$$\frac{1}{T} \int_0^T \mathbb{E}\, \|D_{\theta^*}(\tau, X_\tau, Y_t) - Z\|_2^2\, dt \le \varepsilon_{\text{score}}(\tau).$$

Assumption 3.4 (Lipschitz denoiser)**.** For any $\tau > 0$, there is $L_D(\tau) > 0$, such that for every $x, x', y \in \mathbb{R}^D$,

$$\|D_{\theta^*}(\tau, x, y) - D_{\theta^*}(\tau, x', y)\|_2 \le L_D(\tau) \|x - x'\|_2.$$

Some comments on our assumptions are pertinent. Assumption 3.1 is mild in the sense that for (9) to have strong solutions, Lipschitzness of μ is the usual sufficient assumption in SDE theory. At least local Lipschitzness is necessary. See the discussion in Section 5.6 of Van Handel (2007).

In the same way, under Assumption 3.1 and standard linear growth conditions ensuring existence and uniqueness of strong solutions to (9), Assumption 3.2 follows from Gronwall's inequality (Øksendal, 2003). Observe that Assumptions 3.1 and 3.2 together imply $\int_0^{T+\Delta} \mathfrak{m}_\mu(t)dt < \infty$.

Assumption 3.3 imposes oracle denoiser accuracy, which is standard in diffusion model theory (Gao et al., 2025a; Tang & Zhao, 2025). On the one hand, this avoids the difficult question of obtaining approximation rates for score-matching under specific architectures. At the same time, it also serves as a plug-in assumption that can accommodate any of those rates, depending on the specific architecture chosen by the user.

Finally, Assumption 3.4 holds for many standard neural network architectures, such as any finite MLP with globally Lipschitz activations and fixed weights. In particular, it is satisfied by the architecture described in Appendix B of Tapia Costa et al. (2026), since the x-dependence enters the denoiser through the convolutional and conditioning parts, which are both a composition of linear layers and Lipschitz activations. However, the resulting constant $L_D(\tau)$ is not explicitly controlled during training, and in practice may depend on the dimension D. Regularity assumptions such as 3.4 are also common when analyzing diffusion models (Chen et al., 2023; Gao et al., 2025a; Pfarr et al., 2026).

We are now ready to state our main result, which decomposes the average MSE of the drift estimator into four sources of error.

3.1 The error bound

Theorem 3.5 (Main Theorem)**.** *Let us have a training dataset $\mathcal{D}_{obs}$ as in (10) for certain frequency $\Delta > 0$. Fix the noise level $\tau > 0$ and the number of noisy samples K. Under assumptions 3.1, 3.2, 3.3, and 3.4, we have*

$$\frac{1}{T}\int_0^T \mathbb{E}\big\|\mu(Y_t) - \bar{\mu}_K(\tau, Y_t)\big\|_2^2 dt \le 4\Bigg[\underbrace{L^2\left(\frac{\Delta^2}{T}\int_0^{T+\Delta}\mathfrak{m}_\mu(s)ds + \sigma^2 D\Delta\right)}_{\text{(I) discretization error}} + \underbrace{b_\Delta(\tau)^2\varepsilon_{\text{score}}(\tau)}_{\text{(II) score approximation error}}$$
$$+ \underbrace{L_{\hat{\mu}}(\tau)^2\left(2\alpha_\tau^2\left(\frac{\Delta^2}{T}\int_0^{T+\Delta}\mathfrak{m}_\mu(s)ds + \sigma^2 D\Delta\right) + (1-\sigma_\tau)^2 D\right)}_{\text{(III) noise initialization error}} \tag{23}$$
$$+ \underbrace{\frac{D}{K}L_{\hat{\mu}}(\tau)^2}_{\text{(IV) Monte Carlo sampling error}}\Bigg],$$

where $L_{\hat{\mu}}(\tau) := |a(\tau)| + b_\Delta(\tau)L_D(\tau)$.

Proof. The basic idea, common in the diffusion model literature, is to decompose the error $\|\mu(y) - \bar{\mu}_K(\tau, y)\|_2^2$ into different sources and then bound each term separately. We first introduce the key intermediate quantities.

For any $y \in \mathbb{R}^D$, we define the conditional mean of the single-sample estimator

$$\widetilde{\mu}(\tau, y) := \mathbb{E}_{X_\tau|y}\big[\hat{\mu}(\tau, X_\tau, y)\big]. \tag{24}$$

We also introduce the average single-sample estimator for the case where $X_\tau \sim \mathcal{N}(0, I_D)$. This will capture the noise initialization error:

$$\widetilde{\mu}_{\mathcal{N}}(\tau, y) := \mathbb{E}_{X_\tau \sim \mathcal{N}(0, I_D)}\big[\hat{\mu}(\tau, X_\tau, y)\big]. \tag{25}$$

Finally, we define

$$\mu_\Delta(y) := \frac{1}{\Delta}\,\mathbb{E}\big[Y_{t+\Delta} - Y_t \;\big|\; Y_t = y\big] = \frac{1}{\Delta}\int_0^\Delta \mathbb{E}\big[\mu(Y_{t+h}) \;\big|\; Y_t = y\big]\,dh. \tag{26}$$

This is the time-averaged expected drift over one step starting from y: it only depends on the SDE and the observation frequency Δ. Note that the second equality arises by solving the SDE in $[t, t+\Delta)$.

For each $y \in \mathbb{R}^D$, consider the following decomposition of the error term:

$$\mu(y) - \bar{\mu}_K(\tau, y) = \underbrace{\big[\mu(y) - \mu_\Delta(y)\big]}_{\text{EM discretization}} + \underbrace{\big[\mu_\Delta(y) - \widetilde{\mu}(\tau, y)\big]}_{\text{score approximation}} + \underbrace{\big[\widetilde{\mu}(\tau, y) - \widetilde{\mu}_{\mathcal{N}}(\tau, y)\big]}_{\text{noise initialization}} + \underbrace{\big[\widetilde{\mu}_{\mathcal{N}}(\tau, y) - \bar{\mu}_K(\tau, y)\big]}_{\text{sampling}}. \tag{27}$$

We now proceed to control each term separately.

Term (I): EM discretization error. By definition:

$$\mu(Y_t) - \mu_\Delta(Y_t) = \frac{1}{\Delta}\int_0^\Delta \Big(\mu(Y_t) - \mathbb{E}\big[\mu(Y_{t+h}) \;\big|\; Y_t\big]\Big)\,dh. \tag{28}$$

Taking the squared norm, applying Jensen's inequality and taking expectations, we have

$$\begin{aligned}\mathbb{E}\big\|\mu(Y_t) - \mu_\Delta(Y_t)\big\|_2^2 &\le \frac{1}{\Delta}\int_0^\Delta \mathbb{E}\left[\|\mu(Y_t) - \mu(Y_{t+h})\|_2^2\right] dh \\ &\le \frac{L^2}{\Delta}\int_0^\Delta \mathbb{E}\left[\|Y_{t+h} - Y_t\|_2^2\right] dh.\end{aligned} \tag{29}$$

The first inequality is due to Fubini and the tower property of conditional expectations, while in the second we simply applied Assumption 3.1.

Furthermore, for $h \in [0, \Delta]$, we have

$$Y_{t+h} - Y_t = \int_t^{t+h} \mu(Y_s)\,ds + \sigma(B_{t+h} - B_t), \tag{30}$$

which is the solution to the SDE on $[t, t+h)$. This implies

$$\begin{aligned}\mathbb{E}\big[\|Y_{t+h} - Y_t\|_2^2\big] &\le 2\,\mathbb{E}\left[\left\|\int_t^{t+h} \mu(Y_s)\,ds\right\|_2^2\right] + 2\sigma^2\,\mathbb{E}\|B_{t+h} - B_t\|_2^2 \\ &\le 2h\int_t^{t+h} \mathbb{E}\big[\|\mu(Y_s)\|_2^2\big]\,ds + 2\sigma^2 Dh \\ &= 2h\int_t^{t+h} \mathfrak{m}_\mu(s)ds + 2\sigma^2 Dh.\end{aligned} \tag{31}$$

Observe that in the first inequality we used $(a+b)^2 \le 2a^2 + 2b^2$. For the second inequality we used the Cauchy–Schwarz inequality for the inner product $\langle u, v\rangle := \int_t^{t+h} u(s)\cdot v(s)\,ds$ with $u(s) = 1$ and $v(s) = \mu(Y_s)$, along with the fact that $B_{t+h} - B_t$ is a centered Gaussian with variance h. For the last bound we simply invoke Assumption 3.2 to guarantee that the integral is finite.

Finally, substituting (31) back into (29) and integrating:

$$\begin{aligned}\frac{1}{T}\int_0^T \mathbb{E}\big\|\mu(Y_t) - \mu_\Delta(Y_t)\big\|_2^2 dt &\le \frac{L^2}{\Delta T}\int_0^T\int_0^\Delta \left(2h\int_t^{t+h} \mathfrak{m}_\mu(s)ds + 2\sigma^2 Dh\right) dh\,dt \\ &\le L^2\left(\frac{\Delta^2}{T}\int_0^{T+\Delta} \mathfrak{m}_\mu(s)ds + \sigma^2 D\Delta\right),\end{aligned} \tag{32}$$

where we used Fubini's theorem to swap the integrals.

Term (II): score approximation bias. Recall that $X_\tau = \alpha_\tau Z + \sigma_\tau \varepsilon$, with $\varepsilon \sim N(0, I_D)$. We introduce the *oracle estimator*

$$\widehat{\mu}^{\mathrm{orc}}(\tau, x, y) := a(\tau)x + b_\Delta(\tau)\, \mathbb{E}_{p_0}[Z \mid X_\tau = x,\, Y = y], \tag{33}$$

which replaces the learned denoiser D_{θ^*} by the true conditional expectation. A direct computation using the identity $a(\tau)\alpha_\tau + b_\Delta(\tau) = 1/\Delta$ yields

$$\mu_\Delta(y) = \mathbb{E}\big[\widehat{\mu}^{\mathrm{orc}}(\tau, X_\tau, y) \,\big|\, Y = y\big]\,. \tag{34}$$

Using (34), the gap between μ_Δ and $\widetilde{\mu}$ becomes a conditional expectation of the denoiser error:

$$\begin{aligned}\mu_\Delta(y) - \widetilde{\mu}(\tau, y) &= \mathbb{E}\left[\widehat{\mu}^{\mathrm{orc}}(\tau, X_\tau, y) - \widehat{\mu}(\tau, X_\tau, y) \,\middle|\, Y = y\right] \\ &= b_\Delta(\tau)\, \mathbb{E}\big[\mathbb{E}_{p_0}[Z \mid X_\tau, Y] - D_{\theta^*}(\tau, X_\tau, Y) \,\big|\, Y = y\big]\,. \end{aligned} \tag{35}$$

Taking squared norms, applying Jensen's inequality, and then taking expectation with respect to Y gives

$$\mathbb{E}\|\mu_\Delta(Y) - \widetilde{\mu}(\tau, Y)\|_2^2 \le b_\Delta(\tau)^2\, \mathbb{E}\|\mathbb{E}_{p_0}[Z \mid X_\tau, Y] - D_{\theta^*}(\tau, X_\tau, Y)\|_2^2\,. \tag{36}$$

Since $\mathbb{E}_{p_0}[Z \mid X_\tau, Y]$ is the L^2-optimal predictor of Z among all $\sigma(X_\tau, Y)$-measurable functions, we have

$$\mathbb{E}\|\mathbb{E}_{p_0}[Z \mid X_\tau, Y] - D_{\theta^*}(\tau, X_\tau, Y)\|_2^2 \le \mathbb{E}\|Z - D_{\theta^*}(\tau, X_\tau, Y)\|_2^2. \tag{37}$$

Combining (36) and (37) with Assumption 3.3 yields

$$\frac{1}{T}\int_0^T \mathbb{E}\|\mu_\Delta(Y_t) - \widetilde{\mu}(\tau, Y_t)\|_2^2 dt \le b_\Delta(\tau)^2\, \varepsilon_{\mathrm{score}}(\tau). \tag{38}$$

Term (III): Noise initialization error. Recall the definitions

$$\widetilde{\mu}(\tau, y) := \mathbb{E}_{X_\tau \sim p_\tau(\cdot|y)}[\widehat{\mu}(\tau, X_\tau, y)], \qquad \widetilde{\mu}_{\mathcal{N}}(\tau, y) := \mathbb{E}_{G \sim \mathcal{N}(0, I_D)}[\widehat{\mu}(\tau, G, y)], \tag{39}$$

where $\widehat{\mu}(\tau, x, y) = a(\tau)x + b_\Delta(\tau) D_{\theta^*}(\tau, x, y)$.

By Assumption 3.4, the map $\widehat{\mu}(\tau, \cdot, y)$ is Lipschitz with constant $L_{\hat{\mu}}(\tau) := |a(\tau)| + b_\Delta(\tau) L_D(\tau)$. The Kantorovich–Rubinstein duality (Theorem 5.10 in Villani (2009)) combined with the standard bound $W_1 \le W_2$ then gives

$$\|\widetilde{\mu}(\tau, y) - \widetilde{\mu}_{\mathcal{N}}(\tau, y)\|_2 \le L_{\hat{\mu}}(\tau)\, W_2(p_\tau(\cdot \mid y), \mathcal{N}(0, I_D))\,, \tag{40}$$

where W_p is the $p-$Wasserstein distance. To control the $2-$Wasserstein distance on the right-hand side, we couple $X_\tau \sim p_\tau(\cdot \mid y)$ and $G \sim \mathcal{N}(0, I_D)$ by writing $X_\tau = \alpha_\tau Z + \sigma_\tau \varepsilon$ and $G = \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_D)$, yielding

$$\begin{aligned} W_2^2(p_\tau(\cdot \mid y), \mathcal{N}(0, I_D)) &\le \mathbb{E}\big[\|X_\tau - G\|_2^2 \mid Y = y\big] \\ &= \alpha_\tau^2\, \mathbb{E}[\|Z\|_2^2 \mid Y = y] + (1 - \sigma_\tau)^2 D. \end{aligned} \tag{41}$$

Combining equations (40) and (41), squaring, and taking outer expectation gives

$$\mathbb{E}\|\widetilde{\mu}(\tau, Y_t) - \widetilde{\mu}_N(\tau, Y_t)\|_2^2 \le L_{\hat{\mu}}(\tau)^2 \big(\alpha_\tau^2\, \mathbb{E}\|Z\|_2^2 + (1 - \sigma_\tau)^2 D\big)\,. \tag{42}$$

Applying the increment bound (31) to control $\mathbb{E}\|Z\|_2^2$, and averaging over $t \in [0, T]$, we arrive at

$$\frac{1}{T}\int_0^T \mathbb{E}\|\widetilde{\mu}(\tau, Y_t) - \widetilde{\mu}_N(\tau, Y_t)\|_2^2\, dt \le L_{\hat{\mu}}(\tau)^2 \left[2\alpha_\tau^2 \left(\frac{\Delta^2}{T}\int_0^{T+\Delta} \mathfrak{m}_\mu(s)\, ds + \sigma^2 D\Delta\right) + (1 - \sigma_\tau)^2 D\right]. \tag{43}$$

Term (IV): Monte Carlo error. Recall that

$$\bar{\mu}_K(\tau, y) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mu}(\tau, G^{(k)}, y), \tag{44}$$

with $G^{(1)}, \dots, G^{(K)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_D)$.

Conditionally on $Y = y$, the random vectors $\widehat{\mu}(\tau, G^{(1)}, y), \dots, \widehat{\mu}(\tau, G^{(K)}, y)$ are i.i.d. with mean $\widetilde{\mu}_{\mathcal{N}}(\tau, y)$, so

$$\mathbb{E}_G \left[\|\widetilde{\mu}_{\mathcal{N}}(\tau, y) - \bar{\mu}_K(\tau, y)\|_2^2 \right] = \frac{1}{K} \operatorname{Var}_G\big(\widehat{\mu}(\tau, G, y)\big). \tag{45}$$

Since $\widehat{\mu}(\tau, \cdot, y)$ is $L_{\hat{\mu}}(\tau)$-Lipschitz, applying the Gaussian Poincaré inequality (Boucheron et al., 2013) componentwise yields

$$\operatorname{Var}_G\big(\widehat{\mu}(\tau, G, y)\big) \le \mathbb{E}_G \|J_G \widehat{\mu}(\tau, G, y)\|_F^2 \le D\, L_{\hat{\mu}}(\tau)^2, \tag{46}$$

where $J_G \widehat{\mu}$ denotes the Jacobian with respect to G. Combining (45) and (46) gives

$$\mathbb{E}_G \big[\|\widetilde{\mu}_{\mathcal{N}}(\tau, y) - \bar{\mu}_K(\tau, y)\|_2^2 \big] \le \frac{D}{K} L_{\hat{\mu}}(\tau)^2 = \frac{D}{K} \big(|a(\tau)| + b_\Delta(\tau) L_D(\tau) \big)^2. \tag{47}$$

The proof concludes by applying the inequality $(a+b+c+d)^2 \le 4a^2 + 4b^2 + 4c^2 + 4d^2$ to the decomposition of the error and combining all the bounds.

□

3.2 Interpreting the bound

Theorem 3.5 makes explicit how the different hyperparameters in the estimator (21) interact. We now describe each term separately and discuss their interactions.

- **Term (I): discretization error.** This is the Euler–Maruyama bias coming from replacing the true increment law by its one-step Gaussian approximation. It is of order $O(D\Delta)$, and, as expected, vanishes linearly as the observation grid becomes finer.
- **Term (II): score approximation error.** This term measures how accurately the learned denoiser approximates the expected increments. Since $b_\Delta(\tau) = O(1/\Delta)$ for fixed τ, its contribution is of order

  $$O\left(\frac{\varepsilon_{\text{score}}(\tau)}{\Delta^2} \right).$$

  Thus, for fixed noise level, the effect of denoiser error is upscaled as $\Delta \to 0$. This might come as a surprise at first, but it is natural since the drift contribution to the SDE vanishes faster than the noise as $\Delta \to 0$. Note that the nonlinear relation between the estimator performance and the frequency is also observed empirically in Appendix H of Tapia Costa et al. (2026).
- **Term (III): noise initialization error.** This term quantifies the mismatch between the actual forward marginal $p_\tau(\cdot \mid y)$ and the Gaussian initialization $\mathcal{N}(0, I_D)$. For most VP schedules, such as the one used by Tapia Costa et al. (2026), it vanishes exponentially fast when $\tau \to \infty$ due to the contribution of α_τ and $1 - \sigma_\tau$.
- **Term (IV): Monte Carlo sampling error.** This is the variance introduced by approximating the Gaussian expectation with K i.i.d. samples. For fixed τ and Δ, it decays as $O(D/K)$, as expected.

In summary, decreasing Δ reduces the discretization error in Term (I) but can amplify the score approximation error in Term (II), and also in Terms (III), (IV), via $b_\Delta(\tau)$, so refining the observation grid only helps if the denoiser improves accordingly. Likewise, increasing τ suppresses the initialization error in Term (III) but may worsen the score approximation.

Overall, the bound might not be tight, especially because we are ignoring the dependence of $\epsilon_{\text{score}}(\tau)$ on Δ, and the Lipschitz constant of D_{θ^*} might be large. Nonetheless, it is remarkable that it still captures the nonlinear dependency of the MSE on Δ observed by Tapia Costa et al. (2026).

Furthermore, Theorem 3.5 also allows to choose the hyperparameters $\varepsilon_{\text{score}}$, K and τ depending on Δ, in the same fashion as in Corollary 2.7 of Pfarr et al. (2026):

Corollary 3.6. *Assume $\Delta < 1$ and let $\alpha \in (0, 1]$. Assume moreover that $L_D(\tau)$ remains uniformly bounded along the chosen noise levels. To obtain a bound of rate $\mathcal{O}(\Delta^\alpha)$ in Theorem 3.5, it is enough to choose*

$$\tau \asymp \alpha \log(1/\Delta), \quad K \asymp \Delta^{-(2+\alpha)}, \quad \varepsilon_{\text{score}} \asymp \Delta^{2+\alpha}. \tag{48}$$

Proof. The result follows directly from Theorem 3.5 after basic algebraic manipulations. Observe that we are implicitly assuming that we work with standard VP schedules such as the one used by Tapia Costa et al. (2026): $\beta(\tau) = \gamma_0 + \tau(\gamma_1 - \gamma_0)$, for some $\gamma_1 > \gamma_0 > 0$. □

For example, for a bound of order $\mathcal{O}\left(\sqrt{\Delta}\right)$, we have $\tau \asymp \frac{1}{2}\log(1/\Delta)$, $K \asymp \Delta^{-5/2}$ and $\epsilon_{\text{score}} \asymp \Delta^{5/2}$. We fixed Δ in Corollary 3.6 because the sampling frequency is usually given in many problems, however similar relations could be obtained by fixing any other hyperparameter.

As suggested before, among the requirements in Corollary 3.6, the denoiser accuracy, $\varepsilon_{\text{score}} \asymp \Delta^{2+\alpha}$, is the most stringent. This dependence reflects the $\mathcal{O}(1/\Delta^2)$ amplification factor of $b_\Delta(\tau)^2$ inherent to the estimator. However, the requirement in Corollary 3.6 is very likely overly pessimistic, because Theorem 3.5 treats $\varepsilon_{\text{score}}(\tau)$ as independent of Δ, while in practice finer discretizations increase the training sample size and may yield increments closer to a Gaussian, making the denoising problem easier. Obtaining error bounds for denoiser estimation that take into account the discretization is thus an important avenue for future work.

4 Conclusions and future work

To our knowledge, Theorem 3.5 gives the first theoretical guarantees for SDE parameter estimation via conditional diffusion models, in particular, for the drift estimator of Tapia Costa et al. (2026). The resulting error bound isolates four fundamental sources of error in this approach: discretization bias, denoiser or score-approximation error, noise initialization error and Monte Carlo sampling error.

There are several natural directions for future research. First, our results are based on an oracle assumption on the denoiser or score error. As mentioned above, end-to-end bounds that account for the statistical properties of the underlying neural architecture remain an important open problem. Second, the structure of Theorem 3.5 and Corollary 3.6 suggest that they may be used to obtain practical rules for hyperparameter tuning; we leave this direction for future work.

Furthermore, it would be valuable to extend the present framework beyond the basic setting considered here, and to develop comparable methods and theoretical guarantees for broader families of SDEs, such as equations with time-dependent drifts or unknown diffusion parameter. Finally, it seems possible to extend our theoretical guarantees to SDE sampling methods that also use conditional diffusion models such as Liu et al. (2025); Gao et al. (2025b).

References

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. *International Conference on Learning Representations*, 11, 2023.

Fabienne Comte and Valentine Genon-Catalot. Nonparametric drift estimation for i.i.d. paths of stochastic differential equations. *The Annals of Statistics*, 48(6):3336–3365, 2020.

Fabienne Comte, Valentine Genon-Catalot, and Yves Rozenholc. Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli*, 13(2):514 – 543, 2007.

Christophe Denis, Charlotte Dion-Blanc, and Miguel Martinez. A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation. *Bernoulli*, 27(4):2675–2713, 2021.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025a.

Xuefeng Gao, Jiale Zha, and Xunyu Zhou. Data-driven generative simulation of SDEs using diffusion models. In *NeurIPS Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*, 2025b.

Vincent Guan, Joseph Janssen, Hossein Rahmani, Andrew Warren, Stephen Zhang, Elina Robeva, and Geoffrey Schiebinger. Identifying drift, diffusion, and causal structure from temporal snapshots. *arXiv preprint arXiv:2410.22729*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Peter E. Kloeden and Eckhard Platen. *Numerical Solutions of Stochastic Differential Equations*, volume 23 of *Stochastic Modelling and Applied Probability*. Springer, 1992.

Yury A Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. Springer, 2004.

Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, Geoffrey Schiebinger, et al. Toward a mathematical theory of trajectory inference. *The Annals of Applied Probability*, 34(1A):428–500, 2024.

Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, 2016.

Yanfang Liu, Yuan Chen, Dongbin Xiu, and Guannan Zhang. A training-free conditional diffusion model for learning stochastic dynamical systems. *SIAM Journal on Scientific Computing*, 47(5):1144–1171, 2025.

Sokhna Diarra Mbacke and Omar Rivasplata. A note on the convergence of denoising diffusion probabilistic models. *Transactions on Machine Learning Research*, 2024.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, pp. 25858–25889. PMLR, 2023.

Akihiro Oga and Yuta Koike. Drift estimation for a multi-dimensional diffusion process using deep neural networks. *Stochastic Processes and their Applications*, 170:104240, 2024.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, 2003.

Emanuel Pfarr, Radu Timofte, and Frank Werner. Analyzing the error of generative diffusion models: From Euler-Maruyama to higher-order schemes. *arXiv preprint arXiv:2601.18425*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 9, 2021.

Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations. *Statistic Surveys*, 19:28–64, 2025.

Marcos Tapia Costa, Nikolas Kantas, and George Deligiannidis. Drift estimation for stochastic differential equations with denoising diffusion models. *arXiv preprint arXiv:2602.17830*, 2026.

Ramon Van Handel. Stochastic calculus, filtering, and stochastic control. *Lecture notes*, 2007. URL https://web.math.princeton.edu/~rvan/acm217/ACM217.pdf.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Yuzhen Zhao, Yating Liu, and Marc Hoffmann. Drift estimation for diffusion processes using neural networks based on discretely observed independent paths. *arXiv preprint arXiv:2511.11161*, 2025.

Zheng Zhao, Filip Tronarp, Roland Hostettler, and Simo Särkkä. State-space Gaussian process for drift estimation in stochastic differential equations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5295–5299. IEEE, 2020.